社会化标注中用户动态标签云构建研究*

谢梦瑶 潘旭伟
(浙江理工大学经济管理学院 杭州 310018)

摘要:【目的】标签云可用于信息检索推荐和导航, 由于用户标注具有时序特征, 为有效揭示用户兴趣动态变化, 提出基于时序演化的用户动态标签云构建方法。【方法】利用心理学中记忆的遗忘和加强特征构建标签的动态权重, 从而建立用户动态标签云以反映用户关注点的变化。【结果】与现有的标签云算法比较, 构建的用户动态标签云算法能够根据用户动态变化的兴趣有效地对标签进行排序, 在用户兴趣标签的预测效果上明显高于其他算法, 并具有更高的推荐准确率。【局限】因为用户兴趣在短时间周期内不会有太大变化, 动态的方法在短时间周期内的预测效果不是很显著, 但在长时间周期表现上更为显著。【结论】基于时序演化的用户动态标签云能有效地把握用户当前的兴趣热点, 提高个性化检索和导航的效果。

关键词: 社会化标注 标签 用户兴趣 动态标签云
分类号: TP311

1 引 言

近年来, Web 2.0 为互联网带来了实质性的变化, 用户在网络环境中的角色也从网络信息的被动接受者转换为网络信息的主动创造者。社会化标注作为 Web 2.0 的一个核心构件, 允许大众用户采用标签的方式对自己感兴趣的资源基于自身理解进行无约束标注, 且所有用户的标注都互为可见[1]。以 Flicker、Delicious 为代表的一系列以大众用户参与为中心的社会化标签站点逐渐发展壮大, 社会化标签站点已成为用户有效获取信息资源的一种新渠道。

标签作为社会化标注系统的载体, 成为 Web 2.0 时代一种重要的信息组织工具。社会化标签具有丰富的信息, 标签云(Tag Cloud)作为具有可视权重的标签集合, 它的出现很好地解决了标签信息的可视化问题, 帮助用户快速从大量标签中获取有价值的信息。目前关于标签云的研究主要有排序算法[2]、个性化推荐[3]及可视化布局[4-5]等, 这些研究与实践有力地推动了标签云应用与理论的发展。同时标签云作为一种新兴的检索推荐技术, 通过标签的可视化属性体现不同标签的重要程度, 可以对用户浏览产生导向作用, 从而将用户的关注点吸引到特定字段或区域。Millen 等[6]研究用户在标签云形势下的查询浏览习惯, 发现社会化标签是提高社会化导航的重要途径。Hassan-Montero 等[5]通过定义标签描述资源的程度、覆盖资源的数量等指标的方式来计算标签的有用性, 并通过聚类算法提高浏览体验。此外, 标签云作为社会化信息的导航接口, 通过可视化形式对标签属性及内容分类后可以对不同用户进行个性化搜索推荐。夏瑀等[2]基于 Wikidata 知识库的结构和内容, 通过构建标签云将信息进行标签化处理, 最终实现信息的检索和页面的排序。

现有标签云主要是根据标签累计被标注的次数为权重, 定量地计算出每个标签的权值, 并使用不同的颜色或字体大小形象直观地实现可视化, 以便于用户检索和浏览。用户所使用的标签在一定程度上能够体现用户的兴趣, 随着时间的变化, 用户的兴趣偏好和关注点将产生变化, 而现有基于标签累计频次构建的标签云并不能很好地反映出这种变化。因此, 如何根

通讯作者: 潘旭伟, ORCID: 0000-0001-5583-4000, E-mail: panxw@zstu.edu.cn。
*本文系国家自然科学基金项目“泛在计算环境中社会化驱动的情境感知个性化信息服务研究”(项目编号: 71471165)的研究成果之一。

据用户在不同时间所使用的标签构建出用户动态的标签云，以揭示用户的兴趣和关注点变化，已成为如何利用标签云更好地支持用户信息检索和导航的一个重要问题。为此，本文将从用户标注的时序特征出发，研究反映用户兴趣动态变化的标签云构建方法。

2　用户动态标签云构建

用户动态标签云的构建需要充分考虑时序信息对用户标签使用的影响。根据心理学的遗忘特征和记忆加强现象，将整个用户标注的过程看作是遗忘过程和重复学习过程。即距离用户标注时间越远的标签重要性越低，重复出现的标签重要性又会强化，通过这样的过程动态地计算不同标签的权重，构建动态标签云，以提高用户浏览体验。

2.1　标签权重的动态更新

用户的兴趣会随着时间不断地发生变化，用户兴趣的改变是一种遗忘现象，根据心理学的遗忘特征和记忆加强现象[7]，具有如下基本特征:

(1) 距离当前时刻越近的兴趣具有更高的权重，且兴趣的权重会随着时间的推移逐渐下降;

(2) 当相同的兴趣重复出现时，会有一个兴趣重复强化的过程，与原有兴趣合并构成新的用户兴趣。

所以对于每个兴趣而言，都有遗忘的过程和重复学习的过程。标签是用户自身态度和兴趣的表达，因此可以利用兴趣记忆的遗忘和加强特征来更新用户标签的动态权重，以反映时间对标签权重的影响，从而支持动态标签云的构建。用权重来衡量用户对标签 t_k 的兴趣程度，那么标签权重 w_{t_k} 也有衰减和强化的变化过程，形成多阶段衰减过程，如图 1 所示。

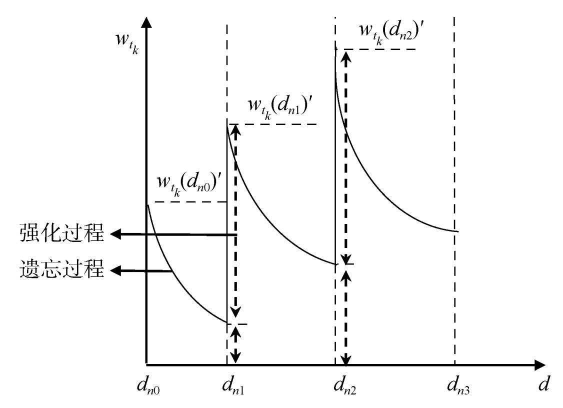


图 1　标签权值 w_{t_k} 多阶段衰减过程

在图 1 中，在某个时间段内(如从 d_{n0} 至 d_{n1})，w_{t_k} 随着时间的推移发生了衰减。而当用户持续地在社会化标签系统中进行标注的过程中，相同的兴趣会阶段性地重复出现(如在 d_{n1} 和 d_{n2} 时刻重复出现了标签 t_k)，则标签权重 w_{t_k} 得到加强而重新上升。这样的重复活动将用户整个标注过程分成多个子阶段，每一个子阶段都是一个新的遗忘过程。因此，根据于洪涛等[7]、印桂生等[8]构建的类似的遗忘曲线计算公式，笔者在此基础上改进并提出标签动态权重的计算公式。标签 t_k 的动态权重 w_{t_k} 的计算涉及的三个主要环节：特定时间点上权重计算、遗忘衰减和记忆加强。

(1) 时间点上的标签权重计算

标签 t_k 在特定时间点 T 上的权重 $w_{t_k}^T$ 通过 TF(词频)方法计算，即用标签 t_k 使用的次数占该时间点上(如可取某一天)所有标签使用次数的比例，计算方法如公式(1)所示。

$$w_{t_k}^T = \frac{tf(t_k)}{\sum_{s=1}^{m} tf(t_s)} \tag{1}$$

其中，m 为该时间点上所有标签数量，$tf(t_s)$ 为标签 t_s 出现的频数。

(2) 标签权重的遗忘衰减

标签 t_k 没有重复出现，那么标签 t_k 的权重 w_{t_k} 随时间发生了衰减，可采用指数形式的遗忘函数进行计算，w_{t_k} 遗忘过程的量化函数定义为公式(2)。

$$w_{t_k}(d) = w_{t_k}(d_{n-1})' e^{-\frac{\ln 2}{hl_u}(d-d_{n-1})} \tag{2}$$

其中，$w_{t_k}(d)$ 为衰减后的标签权重，$w_{t_k}(d_{n-1})'$ 为标签 t_k 第 n–1 次出现时的权重(即上一个遗忘阶段的初始值); hl_u 是用户 u 的半衰期，随着用户知识获取行为周期而不同; d–d_{n-1} 表示距离上次标签 t_k 出现的时间差。

(3) 标签权重的记忆强化

如图 1 所示，在 d_{n1}，d_{n2}，d_{n3} 三个时间点，标签 t_k 重复出现，可以看到 $w_{t_k}(d_n)'$ 的值由上一阶段标签 t_k 权重衰减的剩余量和新的标注活动同一标签 t_k 带来的权重增加量合并而成，使用公式(3)计算每个遗忘阶段的初始兴趣度 $w_{t_k}(d_n)'$。

$$w_{t_k}(d_n)' = w_{t_k}^{new}(d_n) + w_{t_k}(d_{n-1})' e^{-\frac{\ln 2}{hl_u}(d_n-d_{n-1})} \tag{3}$$

其中，$w_{t_k}(d_n)'$ 为 d_n 时间点的初始标签权重，d_n 代表标签 t_k 第 n 次出现的时间点，那么 d_n–d_{n-1} 则是相邻两次标签 t_k 出现的时间差；$w_{t_k}^{new}(d_n)$ 为第 n 次标签 t_k 出现时的权重，其计算方法已经由公式(1)给出，在这里代表在 d_n 时间点的标注活动为标签 t_k 带来的权重增加量。$w_{t_k}(d_{n-1})'e^{-\frac{\ln 2}{hl_u}(d_n-d_{n-1})}$ 即为标签 t_k 权重从上一阶段衰减到 d_n 时间点的剩余量。

2.2　动态标签云构建算法

根据上述标签权重的动态更新机制，建立如下动态标签云构建算法。

输入：用户 u 的标注历史记录(含标注的时间、资源和使用的标签)

输出：用户 u 的动态标签云

算法描述：

①利用公式(1)计算用户标注初始的标签兴趣权重，得到每个标签在不同时间点上(通常以天为计)的权重。

②将标签按时间先后顺序进行排序，判断标签 t_k 有没有重复出现，如果没有则进入步骤③对标签权重进行更新；如果有则进入步骤④对标签权重进行更新。

③根据公式(2)计算标签衰减后的标签权重。

④根据公式(3)计算强化后的标签权重，由上一阶段的衰减值和新标注活动带来的权重增加量合并而成。

⑤综合每个标签的权重并进行归一化处理，得到用户 u 的动态标签云。

3　实验研究

3.1　实验数据

实验数据来源于 Last.fm 和 Delicious 两个具有代表性的社会化标注系统，Delicious 的数据取自北京大学 DAIM 研究组收集的 Delicious 网站在 2009 年 1 月至 6 月期间 18 万多用户的社会化标注数据，下载网址为：http://www.datatang.com/data/42989；Last.fm 的数据取自于马德里自治大学信息检索组收集的 1 892 名用户的音乐标注数据信息，下载网址为：http://grouplens.org/datasets/hetrec-2011。实验数据的基本统计如表 1 所示。

表 1　实验数据基本统计

数据集	用户数	资源数	标签数	时间跨度
Delicious	185 068	4 153 293	939 036	2009.1-2009.6
Last.fm	1 892	17 632	11 946	2005.8-2011.5

本文选取活跃度(即标注的资源数)较高的用户，对其标注活动数据进行实验研究，选取的标注数据包含用户、资源、标签和标注时间等信息，若有多个标签用于一个资源的标注，则形成多条记录，示例数据如表 2 所示。

表 2　标注数据示例

UserID	ResourceID	TagID	TagTime
626177	7864044	2521	2009/1/27
626177	7864044	7833	2009/1/27
626177	7864044	7192	2009/1/27
626177	34862262	7833	2009/1/29
626177	34862262	94	2009/1/29
625254	7864044	2521	2009/1/30
625254	7864044	7833	2009/1/30
625254	7864044	289	2009/1/30
625254	7864044	5032	2009/1/30

3.2　用户动态标签云的可视化

为直观反映动态标签云的效果，选取典型用户的标签云可视化结果进行比较。分别使用现有的累计标注频次方法和上述提出的动态标签云构建方法建立可视化标签云，以字体大小对标签权重进行区分。图 2 为 Delicious 某活跃用户(UserID: 12116)到第 6 个月末时(即标注截止时间)两种方法构建的标签云可视化结果(对前 50 个热门标签进行显示)，标签的字号越大表示该标签的权重越高。该用户在 6 个月内共进行 2 710 次标注，其中标注资源数 995 个，使用标签数 424 个，使用最多的标签频次达到 447 次，最低的为 1 次。

(a) 现有累计标注频次构建的静态标签云

(b) 考虑标注时序信息的动态标签云

图 2　两种不同方法构建的可视化标签云比较

从图 2 的可视化标签云可见，以现有累计标注频次构建的静态标签云与本文提出的考虑标注时序信息构建的动态标签云所得到的标签权重相对大小是有差异的，因而在引导用户对信息检索和导航的相对优先顺序上也将产生不同效果。

3.3　用户动态标签云导航效果评价

为进一步检验提出构建的动态标签云是否更好地反映了用户兴趣的变化，起到更好的信息检索和导航作用，开展了定量化的比较实验研究。因为若用户当前对某个标签感兴趣，则其会在未来一段时间继续使用，所以以在某个时间点上标签权值大的前 N 个标签在未来一段时间内被用户再次标注使用的情况构建评价指标，以表征标签云对用户兴趣刻画情况。为此首先定义两个基本的评价指标公式，分别如公式(4)和公式(5)所示。

$$Acc = \frac{n_N}{N} \tag{4}$$

$$Rec = \frac{n_N}{n_0} \tag{5}$$

其中，n_N 为排名前 N 个标签在未来一段时间内被标注使用的总次数；n_0 为在未来一段时间内用户标注使用的标签总次数。从指标定义可见，Acc 表示排名前 N 个标签在未来一段时间被标注使用的平均频次，Rec 表示排名前 N 个标签使用的总频次占用户在该段时间内所有使用标签频次的比率。因此，这两个基本指标分别从不同视角表征排名前 N 个标签在未来一段时间被使用的情况，为此结合二者定义新的指标作为综合评价指标，如公式(6)所示。

$$AR = Acc \cdot Rec = \frac{n_N}{N} \cdot \frac{n_N}{n_0} = \frac{n_N^2}{N \cdot n_0} \tag{6}$$

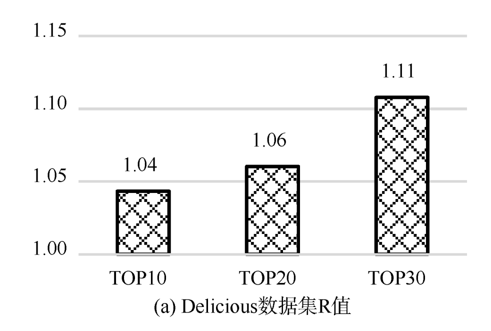
以 Delicious 实验数据集为例，选取数据集中 15 个有比较完整标注历史的用户(即这些用户基本上有 6 个月的持续标注活动)为对象。将数据集跨越的 6 个月时间按先后顺序分为 36 个周期，每个周期为 5 天。令 k=1, 2, …, 36，取第 k 个周期结束时的标签权值，按大小排名前 10、前 20 和前 30 个标签，计算这些标签在第 k+1 个周期的 5 天内的使用情况，计算出 AR 评价指标值。例如，当取 k=2 时，计算出到第 10 天时用户标签权重排名的前 10、20 和 30 个标签，然后使用之后 5 天(即从第 11 天到第 15 天的第 3 个周期)的用户标注数据来计算评价指标值，依次类推。为避免用户在标注开始阶段的随机性影响，在具体实验过程中从第 20 个周期开始计算相应的评价指标值。同样，Last.fm 以 15 个活跃用户为研究对象，由于该数据集时间跨度比较大，将数据集跨越的 6 年时间按先后顺序分为 12 个周期，每个周期 6 个月，评价方法同 Delicious 数据集。

为检验本文提出的标注时序信息构建的动态标签云与现有以累计标注频次构建的静态标签云在反映用户兴趣度上的差异，在分别计算得到两者综合评价指标 AR 值后，计算它们的比值，如公式(7)所示。

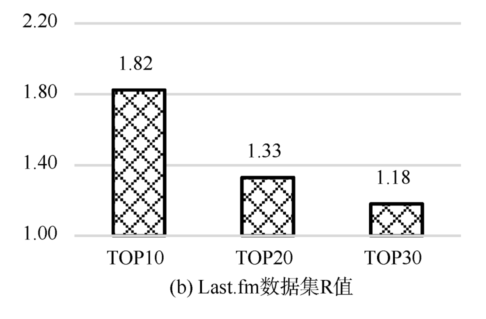
$$R = \frac{AR_D}{AR_S} \tag{7}$$

其中，AR_D 和 AR_S 分别为动态标签云方法和静态标签云方法计算得到的综合评价指标值。

图 3 为选取 15 个用户用两种方法进行比较的结果。可见，在观测的标签权值排前 10、20 和 30 个标签，动态方法相较于静态方法都有不同程度的提升，Delicious 数据集的提升从 4%到 11%不等，而 Last.fm 数据集提升显著，从 18%到 82%。



(a) Delicious数据集R值



(b) Last.fm数据集R值

图 3　动态标签云相较于静态标签云的提升效果

现有的实验结果来看，Delicious 数据集的提高效果并不是十分显著，因为在 6 个月的短时间内的兴趣通常并没有发生本质上的变化，对于时间周期更长的

Last.fm 数据集，动态方法相较于静态方法的提高效果显著，这表明基于标注时序信息构建的动态标签云能够更好地反映用户兴趣的变化，更好地起到对信息检索和导航作用。

另外，在上述Last.fm实验数据和相同的测评指标下，将动态方法和其他方法进行对比实验，选择的对比方法为基于累计频次的方法(TF)和文献[8]提出的标签时间权策略，包括 TF 时间权和 TFIDF 时间权。由图 4 可知，动态方法的效果高于其他方法，具有较高的推荐准确率。

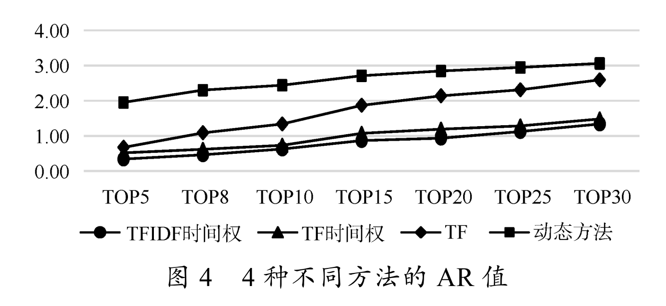


图 4　4 种不同方法的 AR 值

4　结　语

为使标签云更好地反映出用户当前的动态兴趣，基于用户兴趣动态变化的特性和社会化标注的时序特征，提出基于时序演化的用户动态标签云构建方法。该方法基于心理学中记忆的遗忘和加强特征构建标签的动态权重。从实验结果看，动态标签云的可视化结果有别于现有以累计标注频次构建的静态标签云，与现有的标签排序算法比较，动态方法优于其他算法，并且能够刻画和把握用户当前兴趣，有利于更好地帮助用户利用标签云进行信息检索和导航。标签只是反映用户兴趣的一个单一词汇，而用户的兴趣往往是由多个标签形成的集合所刻画的主题，因此如何在现有动态标签云基础上进一步挖掘用户的兴趣主题将是下一步开展的工作。

参考文献:

[1] 魏建良，朱庆华. 基于社会化标注的个性化推荐研究进展[J]. 情报学报，2010，29(4)：625-633. (Wei Jianliang，Zhu Qinghua. Advances in Personalized Information Recommendation Based on Social Tagging [J]. Journal of the China Society for Scientific and Technical Information, 2010, 29(4): 625-633.)

[2] 夏瑀，葛佳琦，马秀，等. 基于 Wikidata 和标签云的搜索算法研究[J]. 软件导刊, 2016, 15(8): 42-46. (Xia Yu, Ge Jiaqi, Ma Xiu, et al. Research on Searching Algorithm Based on Wikidata and Tag Cloud[J]. Software Guide, 2016, 15(8): 42-46.)

[3] 曾子明，张振. 社会化标注系统中基于社区标签云的个性化推荐研究[J]. 情报杂志，2011，30(10)：128-133. (Zeng Ziming，Zhang Zhen. A Personalized Recommendation Approach Based on Community Tag Cloud in Social Tagging System [J]. Journal of Intelligence, 2011, 30(10): 128-133.)

[4] 毕强，周姗姗，马志强，等. 面向知识关联的标签云优化机理研究[J]. 现代图书情报技术，2014(5)：33-40. (Bi Qiang, Zhou Shanshan, Ma Zhiqiang, et al. Study on Optimization Mechanism of Tag Cloud for Knowledge Relation [J]. New Technology of Library and Information Service, 2014(5): 33-40.)

[5] Hassan-Montero Y, Herrero-Solana V. Improving Tag-Clouds as Visual Information Retrieval Interfaces[C]//Proceedings of International Conference on Multidisciplinary Information Sciences and Technologies, Merída, Spain.2006.

[6] Millen D R, Feinberg J. Using Social Tagging to Improve Social Navigation [OL]. [2016-09-30]. https://www.researchgate.net/publication/228904554_Using_social_tagging_to_improve_social_navigation.

[7] 于洪涛，崔瑞飞，董芹芹. 基于遗忘曲线的微博用户兴趣模型[J]. 计算机工程与设计，2014，35(10)：3367-3372, 3379. (Yu Hongtao, Cui Ruifei, Dong Qinqin. Micro-blog User Interest Model Based on Forgetting Curve [J]. Computer Engineering and Design, 2014, 35(10): 3367-3372, 3379.)

[8] 印桂生，崔晓晖，马志强. 遗忘曲线的协同过滤推荐模型[J]. 哈尔滨工程大学学报，2012，33(1)：85-90. (Yin Guisheng, Cui Xiaohui, Ma Zhiqiang. Forgetting Curve-based Collaborative Filtering Recommendation Model [J]. Journal of Harbin Engineering University, 2012, 33(1): 85-90.)

[9] 赵开慧. 基于社会化标注的个性化信息推荐方法研究[J]. 情报科学，2015，33(6)：39-42. (Zhao Kaihui. Recommendation Method of Personalized Information Based on Socialized Tagging [J]. Information Science, 2015, 33(6): 39-42.)

作者贡献声明:

谢梦瑶：方法研究和数据实验，论文起草；
潘旭伟：提出研究思路、实验方案设计和论文修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据[1]见期刊网络版 http://www.infotech.ac.cn; 支撑数据[2-4]由作者自存储, E-mail: panxw@zstu.edu.cn。

[1] 谢梦瑶, 潘旭伟. 实验数据.zip. 选取的 Delicious、Last.fm 的用户实验数据.

[2] 谢梦瑶, 潘旭伟. 静动态方法比较.xlsx. 动态方法相对于静态方法的提升效果.

[3] 谢梦瑶, 潘旭伟. 动态方法与其他时间权方法的比较.xlsx. 动态方法相对于其他方法的提升效果.

[4] 谢梦瑶, 潘旭伟. 实验数据库文件.sql. 数据库代码.

收稿日期: 2016-10-12
收修改稿日期: 2016-12-09

Constructing Dynamic Social Tag Cloud for User Interests

Xie Mengyao   Pan Xuwei
(School of Economics and Management, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: [**Objective**] Social tags can be used for the recommendation and navigation sections of information retrieval systems. This paper proposes a method to construct a dynamic user tag cloud based on the temporal evolution to reveal the changes of user interests. [**Methods**] We established the tags' dynamic weights with the forgetting and strengthening characteristics of memory in psychology. Thus, the dynamic user tag cloud reflect user's changing focus. [**Results**] Compared with the existing ones, the proposed algorithm could effectively sort the tags, and then make accurate predictions or recommendations. [**Limitations**] The proposed method performed well over long period of time because user's interests do not change significantly in a short period of time. [**Conclusions**] The proposed algorithm could effectively identify user's interests and then improve the personalized services.

Keywords: Social Tagging   Tag   User Interests   Dynamic Tag Cloud